Fake News Detection

Divine Mbunga, 17324651
Nissimol Aji, 17321973

Introduction

The project developed is intended to distinguish between fraudulent and real news.The deliberate spread of misinformation has generated an issue of fake news that is more prominent today. Due to the widespread use of technology, news reaches different corners of the world within seconds. This is interesting because news has the potential to divide people around the world especially when they contradict. For example, the argument of whether masks should be worn during this pandemic.

A natural question one could ask is how to draw the line between what does it mean for a news to be fraudulent and real. After observing different aspects that define an article to be real or fake, the team decided to group news articles based on identifying sources that are known world-wide and sources that are not well known.For instance, BBC is a source majority of readers are aware of, while InfoWars was recently banned by the Google app store for false coronavirus conspiracy claims, this news outlet would not be considered reliable compared to BBC.

The input to our algorithm are articles from various news websites such as CNN, BBC, The Guardian, InfoWars, Breitbart, and The Onion . We then use logistic regression and a kNN classifier to output a predicted +1 meaning real news coming from CNN/BBC/The Guardian and -1 meaning fake news coming from InfoWars/Breitbart/TheOnion.

Dataset and Features

WEB SCRAPING

Web Scraping is the extraction of data from websites.For this project, the data that needed to be extracted was articles from different sources.Some of the libraries that were used in the process was BeautifulSoup and NewsPapers 3k. For each news outlet a number of articles were scrapped and stored as a csv format. Once all the articles were gathered all the separate datasets were merged into one large dataset containing 970 articles. A snapshot of the dataset can be seen below.

	A	B	C	D	E	F	G
1	Title	Author	Text	Resource	Label		
2	76 Francis	['Breitbart	BERLIN	BreitBart	-1		
3	Boris Face	['Victoria	Boris	BreitBart	-1		
4	Tucker Car	['Jeff Poor	Monday,	BreitBart	-1		
5	David Mar	['David Ng	Acclaime	BreitBart	-1		
6	Coronavir	['Breitbart	BETHLEH	BreitBart	-1		

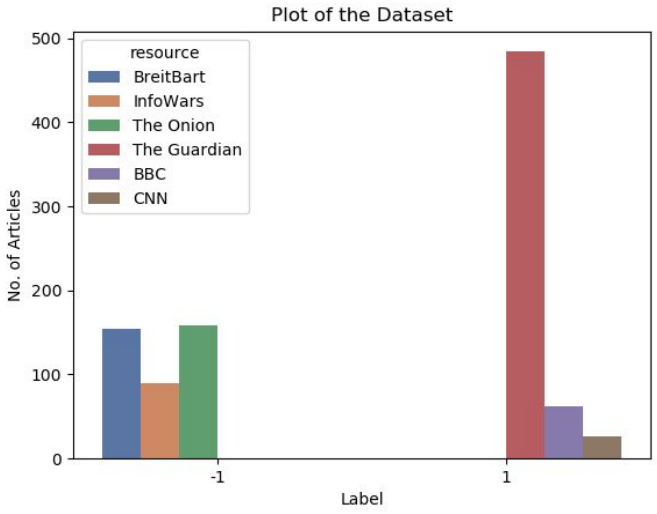
NORMALISATION & PREPROCESSING

The data extracted was preprocessed using the nltk library.The first task was to remove any columns or rows with null values.This was followed by the removal of punctuations,all the characters were converted to lowercase, the stopwords were removed which are words in the text that have less meaning like “the”, “and”, “or” etc. All the sentences were tokenized and a new clean dataset was created with the title,author and text merged to form a new feature called Article. Below is a snapshot of the cleaned data.The data was already normalised therefore no normalization was required.

	Article	Resource	Label
0	76 franciscan nuns from same german monastery ...	BreitBart	-1
1	boris faces massive tory rebellion coronavirus...	BreitBart	-1
2	tucker carlson dr fauci was revealed powermad ...	BreitBart	-1
3	david mamet calls out experts questions lockdo...	BreitBart	-1
4	coronavirus robs biblical bethlehem christmas ...	BreitBart	-1

THE TRAINING DATA

Here is a plot of the dataset, where the y-axis is the number of articles and the x-axis contains two labels, -1 for “fake news” which is BreitBart, InfoWars and The Onion and +1 for “real news” which contains the resources CNN, BBC and the Guardian. We can see from the legend and plot that the articles are in the correct category.



Methods

The news articles are classified as fake or true using logistic regression and KNN classifier. Since machine learning models do not accept the raw text as inputs, the articles are converted into vectors of numbers using count vectorizer and TF-IDF transformer.The count vectorizer firstly tokenizes the sentences thus creates a vocabulary of words for each article. Using this collection of vocabulary, frequencies of each word (term) in the articles are generated using the TF IDF transformer. This will be the input to both models. The sklearn Pipeline library was used to bind the term frequencies,count vectoriser and the model into an interface.A sequence of features(words) were transformed and correlated together in the model for the prediction task.

LOGISTIC REGRESSION

Logistic regression is a classification algorithm used for binary classification and suitable for this project to distinguish between fake or real news.After creating a 80/20 train-test split of the dataset.The logistic regression classifier is applied to the TF-IDF features building the model on the default parameters.The logistic regression classifier uses the weighted combination of the input features i.e the Articles and

passes them through a sigmoid function.The sigmoid function transforms any real number input to a number between 0 and 1.However,since the number of features is higher than the number of data points,the model tends to be underdetermined meaning that vocabulary of words in each article is somewhat unique and thus could produce infinitely many solutions.To fix this issue,hyper parameters are introduced which is later discussed in the next section.

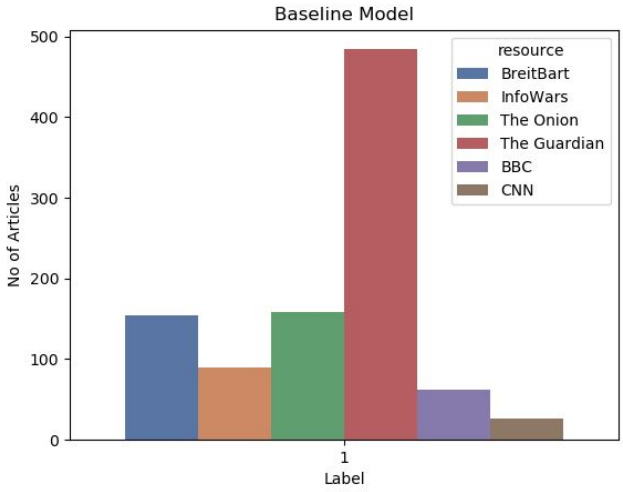
KNN CLASSIFIER

The k-Nearest Neighbour classifier, it directly uses the training data to make predictions. The model calculates the distance between data points and selects k training points that are closest to a point on the plane. For example if k=3 and two articles were under the label -1 and one was under +1 then the model will predict -1. The weighting of the neighbouring articles are uniform to attach equal weight to all the data to allow for majority vote. Increasing k will tend to smooth out the function, causing under-fitting and decreasing k will make the function more complex, which can cause over-fitting. Cross validation is used to determine a suitable value for this classification problem and is discussed in the next section.

Experiments/Results/Discussion

BASELINE MODEL

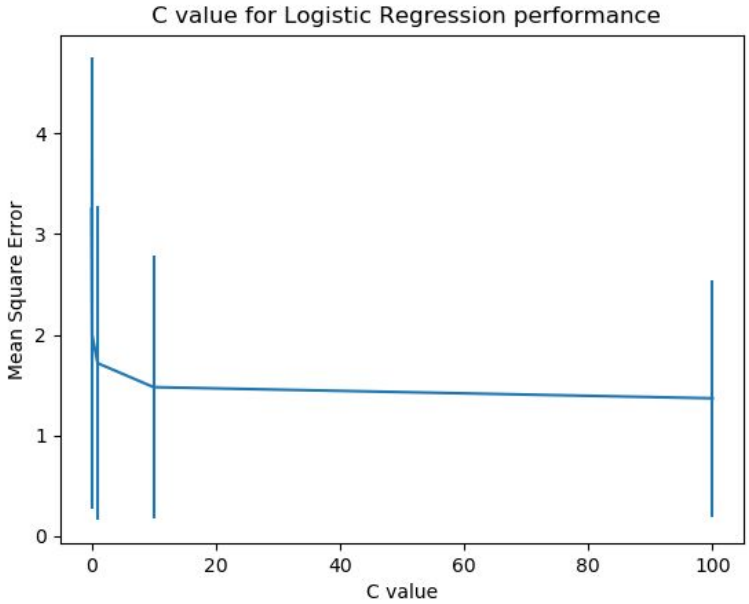
The sklearn DummyClassifier class was used to produce a baseline model for this project where the most common class, -1 or +1, was constantly predicted. Beside is a plot of the predictions made from this model.



LOGISTIC REGRESSION

1. Cross Validation to select Hyperparameter C

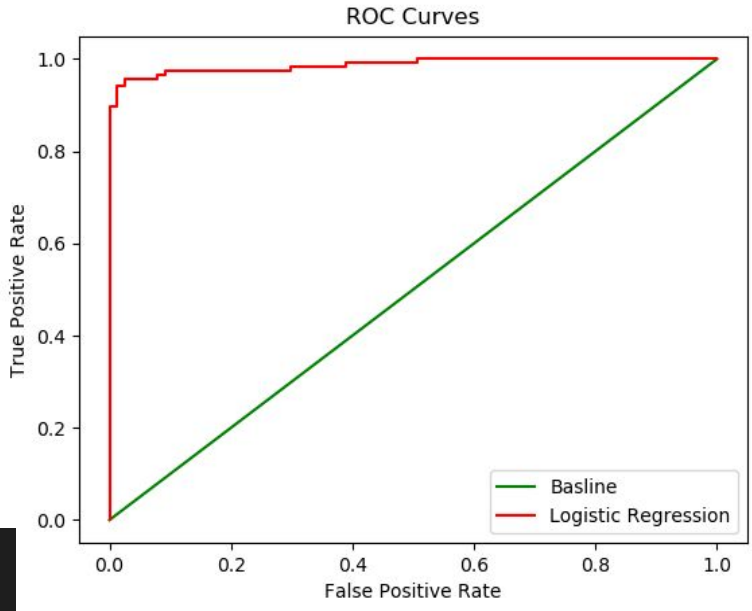
As mentioned above in Methods Section,hyper parameter C is introduced to the model to improve its accuracy.To find which C is suitable for the model cross validation technique is applied.This technique involves dividing the training data into k parts,k-1 parts are used for training and 1 for testing.The procedure is iterated k times each time rotating the test set. Default value for K =5 was chosen.An error plot was



obtained for different ranges of C=[0.001,0.1,1,10,100].These values were chosen to see the range of prediction error when c is small and when c is large.The plot can be seen above.From analysing the plot,its clear when c is small there is a large prediction error. C = 10 was chosen based on looking at the plot there is a small prediction error.The reason for not choosing C= 100 is because of overfitting the data.

2. ROC curve and Confusion Matrix

From observing the ROC plot, The logistic regression model is a good fit of the data. Once it reaches a threshold The curve is close to the point 1 in the plot which means that that the model is accurately predicting true positives and its well off the baseline. This model is making predictions accurately.The confusion matrix clearly sides with this evidence that this is a good model showing 95% accurately predicting correct labels.



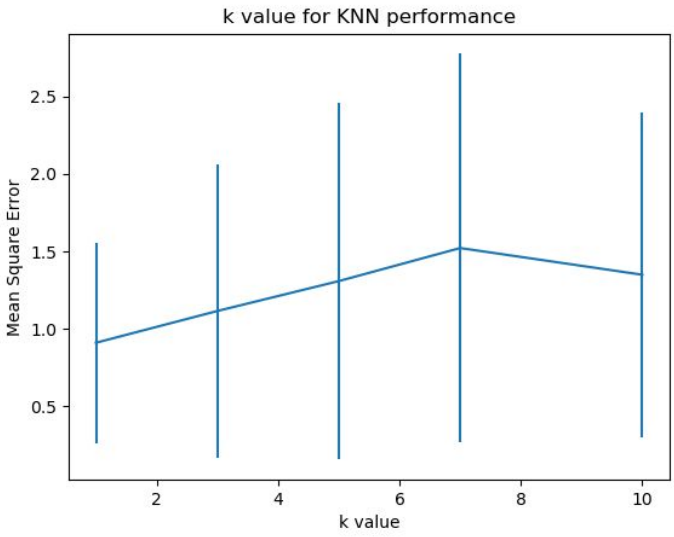

```
Baseline Model
 [[  0  77]
 [  0 118]]
Accuarcy: 0.6051282051282051
Logistic Regression
 [[ 76   1]
 [  7 111]]
Accuracy: 0.958974358974359
```

KNN CLASSIFIER

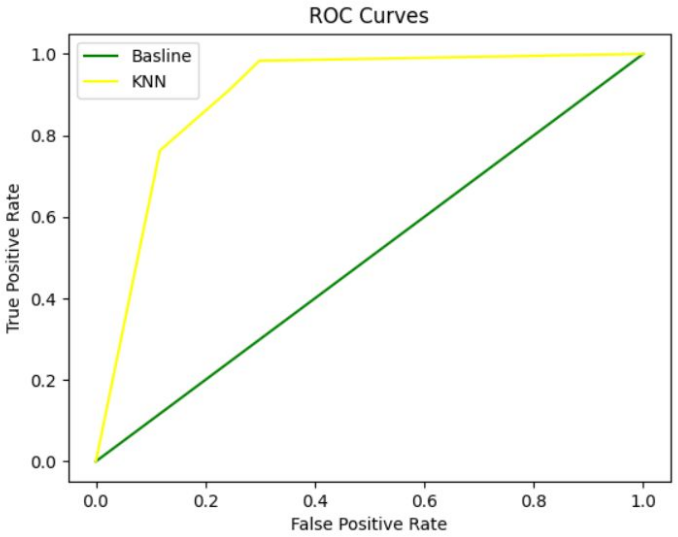
1. **Cross Validation to select Hyperparameter C**

Cross validation was used to select a suitable k value. This was done using K-Folds with 5 folds. The range of k values evaluated were; 1,3,5,7 and 10. Beside is an error plot showing the performance of the k values. Here we can see that taking the variance, error and the precaution of over and under fitting of the data k=3 is suitable for this model.

The data was split into 80% training data and 20% test data, the standard. This was to ensure that the model had enough data to train with and could be evaluated with new data for a better analysis.



2. ROC curve and Confusion Matrix

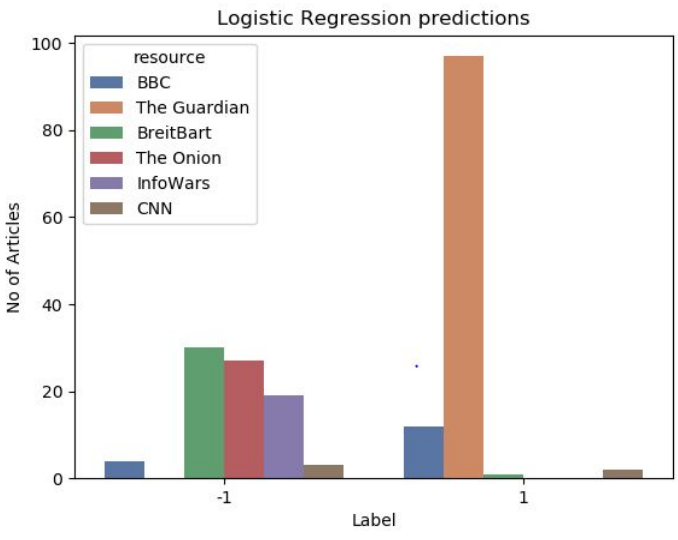


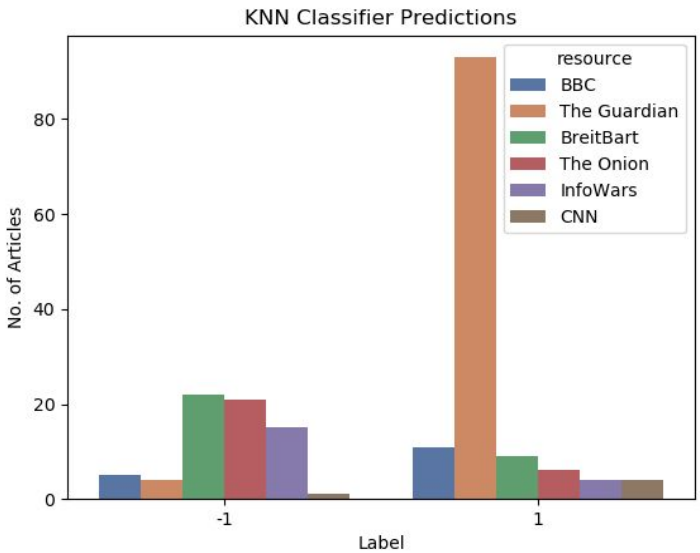
The metrics considered in this model are accuracy, the confusion matrix and an ROC curve comparing this model to the baseline model. The accuracy of this model for the test data is 0.85 which is better than the baseline with an accuracy of 0.61, this is a good sign. From the ROC curve below we can see that the kNN model is doing very well as it is closer to the ideal "perfect" classifier located in the leftmost corner of the plot and is away from the baseline. Also from the confusion matrix below we see that the kNN model has a count of 58 true negatives, 108 true positives with only 10 false negatives and 19 false positives. This is great compared to the baseline model's 77 false positives.With all this information beside is a plot of the predictions that this model made.

```
Baseline Model
 [[  0  77]
 [  0 118]]
Accuarcy: 0.6051282051282051
KNN Classifier
 [[ 58  19]
 [ 10 108]]
Accuracy: 0.8512820512820513
```

Summary

From observing the various models above it's quite clear that Logistic regression performs better than KNN classifier.This conclusion can be deduced from analysing both the ROC curves and confusion matrix for the two models. Logistic regression has an accuracy of 95% while kNN is 85% as seen in the confusion matrix. When observing the ROC curve for logistic regression,the plot curves and bends earlier compared to the KNN classifier. This is because the kNN model only takes the closest three 'neighbours' into consideration to predict a label.However after a few iterations the model learns how to classify between real and fake as it propagates through more neighbours.Whereas the logistic regression model takes the whole training data into consideration and uses that to attempt to distinguish between the two labels. Below from the Logistic regression prediction plot, all articles from The Guardian have been predicted as real, which is correct while the kNN model predicts some of the articles from The Guardian as fake. Only a small number of articles from BreitBart articles were predicted as real when it should be fake in the Logistic model. More articles from BreitBart were predicted as real in the kNN model which is incorrect. In summary, logistic regression performs better than KNN model for this dataset.





Contributions

TASK	Nissimol Aji	Divine Mbunga
Scrape “real” articles: CNN,The Guardian & BBC		✓
Scrape “fake” articles: BreitBart, InfoWars & The Onion	✓	
Merging the Dataset and create sample		✓
Cleaning the dataset	✓	
Logistic Regression Model: ● Cross validation ● Confusion matrix ● ROC curve ● Report:Logistic Regression section in Methods and in Experiments/Results/ Discussion section	✓	
kNN Classifier: ● Cross validation ● Confusion matrix ● ROC curve ● Report: kNN Classifier section in Methods and in Experiments/Results/ Discussion section		✓
Report: ● All of the report except Logistic and kNN both of us wrote it together	✓	✓

SIGNATURE:

Divine Mbunga

Nissimol Aji

Divine Mbunga

Nissimol Aji

Source code link

https://github.com/nissimanjayil/Fake-News-Detection